ABSTRACT OF THE DISCLOSURE

The present invention provides a throttling system that can throttle incoming requests to a server that includes a variable sized buffer for holding incoming calls prior to processing by the server. The number of requests that are held in a queue by the buffer can be dependent on the overload status of the server. If the server is not overloaded, the number of requests that are held in the buffer can be large, such as the full capacity of the buffer. Alternatively, if the server is overloaded for a predetermined amount of time, then the number of requests that are held in the buffer can be decreased, such as to only a portion of the full capacity of the buffer. Any requests that arrive at the buffer once the buffer is at its capacity can be discarded or blocked. Accordingly, a reduction of the buffer size in a overloaded state results in a superior delay performance without increased request blocking of the processor.